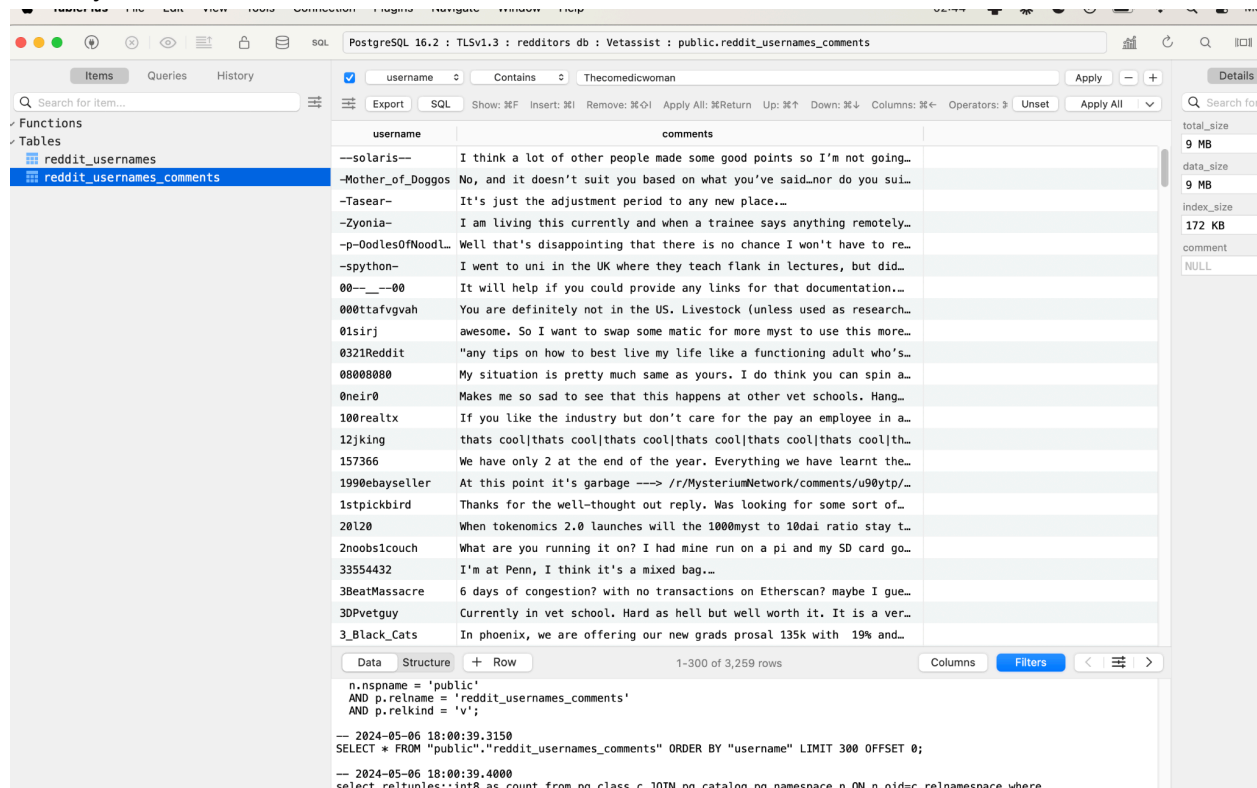


Create a classifier that will accurately classify a list of reddit comments into the proper labels.

You can access the data here in this postgres database

postgresql://niphemi.oyewole:W7bHlgaN1ejh@ep-delicate-river-a5cq94ee-pooler.us-east-2.aws.neon.tech/Vetassist?statusColor=F8F8F8&env=&name=redditors%20db&tLSMode=0&usePrivateKey=false&safeModeLevel=0&advancedSafeModeLevel=0&driverVersion=0&lazyload=false

Inside you will find a table with a list of reddit usernames and their comments.



The screenshot shows a PostgreSQL database interface with a table named 'reddit_usernames_comments'. The table has two columns: 'username' and 'comments'. The interface includes a search bar, a table structure view, and a query editor. The table contains 3,259 rows of data. The query editor shows a query that selects all rows from the table, ordered by username, with a limit of 300 rows and an offset of 0.

username	comments
--solaris--	I think a lot of other people made some good points so I'm not going--
-Mother_of_Doggos	No, and it doesn't suit you based on what you've said...nor do you sui...
-Tasear-	It's just the adjustment period to any new place...
-Zyonia-	I am living this currently and when a trainee says anything remotely...
-p-OodlesOfNoodl...	Well that's disappointing that there is no chance I won't have to re...
-spython-	I went to uni in the UK where they teach flank in lectures, but did...
00--__--00	It will help if you could provide any links for that documentation...
000ttafvgh	You are definitely not in the US. Livestock (unless used as research...
01sirj	awesome. So I want to swap some matic for more myst to use this more...
0321Reddit	"any tips on how to best live my life like a functioning adult who's...
08008080	My situation is pretty much same as yours. I do think you can spin a...
0neir0	Makes me so sad to see that this happens at other vet schools. Hang...
100realtx	If you like the industry but don't care for the pay an employee in a...
12jking	thats cool thats cool thats cool thats cool thats cool thats cool th...
157366	We have only 2 at the end of the year. Everything we have learnt the...
1990ebayseller	At this point it's garbage ----> /r/MysteriumNetwork/comments/u90ytp/_...
1stpickbird	Thanks for the well-thought out reply. Was looking for some sort of...
20120	When tokenomics 2.0 launches will the 1000myst to 10dai ratio stay t...
2noobs1couch	What are you running it on? I had mine run on a pi and my SD card go...
33554432	I'm at Penn, I think it's a mixed bag...
3BeatMassacre	6 days of congestion? with no transactions on Etherscan? maybe I gue...
30Pvetguy	Currently in vet school. Hard as hell but well worth it. It is a ver...
3_Black_Cats	In phoenix, we are offering our new grads prosal 135k with 19% and...

Your classifier should run through this list and determine if they are of these categories:

- A. Medical Doctor
 - a. This label should only include practicing doctors and or consultants to doctors/clinics.
 - b. Medical school students, nurses or medical professionals who aren't doctors should go into the "Other" label (C) instead
- B. Veterinarian
 - a. This label should only include practicing vets and/or consultants to vets/clinics.
 - b. Veterinarian students or veterinarian technicians should go into the "Other" label (C) instead
- C. Other
 - a. Anyone who does not fit within the Medical Doctor, or a Veterinarian label

Training Set

You will need to manually label and/or use an LLM to label some of the data inside the postgres database to provide labeled data for your training set.

Here's an example of what the labels should look like (the "reason" column is unnecessary, but provided in this example here for you to understand "why" a specific label was assigned to a specific comment):

https://docs.google.com/spreadsheets/d/1Kj7217yfvIcJN7WMr7HISsfHgNQ8dwSsX9LF1v0F_tU/edit?usp=sharing

You may take a small portion of the results in the database to form the training set, then use a portion of the database as the evaluation set. You do not need to use all of the data in the database or even many of them, label only as much as required for your classification model to perform well (to see how well your classification model needs to perform see the "**Objective/Criterias**" section below).

LLMs can be used for helping you produce the training data, however, an LLM or language model cannot be used as the classification model itself.

If you want feedback on your training set, you are allowed to send your labeled training set as a csv and we can give you feedback if it is correct and/or fix to send back to you so that you won't run into any issues on the training set.

Objective/Criterias

These are the criterias we will be evaluating you on:

1. How well written is your code?
 - a. Is it properly formatted?
 - b. Easy for other engineers to read and understand?
2. How well did your trained classifier perform?
 - a. Did it properly categorize the data and assign the right labels?
 - i. To evaluate your model's outputs:
 1. we will take 20-30 new reddit comments that are not in the database and label them manually similar to [document](#)
 2. we will run your model on these 20-30 new comments
 3. we will compare our "manual" labels from (1) compared to the labels your model produced in (2)
 4. we expect 70+% to match between our "manual" labels and your model's outputs for you to complete this challenge successfully

- ii. If you did not properly classify to the above threshold (70+%), assume you have failed this challenge. We will not evaluate any candidates who do not complete the classification.
- 3. Did you choose a good classifier model/approach for this task?
 - a. It should be powerful/performant enough to provide the highest quality results to accomplish the task.
 - b. LLMs and language models are not allowed for the classifier model, they can only be used to produce the training set

If you don't get a passing to outstanding grade on all the criterias above (1-3) assume you will not be considered for the role. A poor/mediocre grade on any one of the criterias above will fail your assessment automatically.

We evaluate you 100% on your work and not your past portfolio/resume etc, if you can pass this coding challenge you are very likely to be hired by our team so put your best effort into this challenge as it is the core item we will evaluate you on.

Timelines

You have 1 week to complete this challenge, though this is not a hard requirement. If you require more time that can be allowed.

Although we give 1+ week of time, workload-wise we expect this assessment to take no more than a few hours of work if you choose an appropriate model and training process. Our team member who made this coding challenge, completed this within that time frame while preparing this coding challenge so that is our expectation.

When you are done please send us the link to your code (can be any format: github repo link, colab link, etc) and a csv with the columns (username, comment, label).

Other

If you have any questions please email us. If you require GPU or other resources for training your model, we may be able to provide a stipend to pay for any expenses incurred up to a reasonable amount, email us if you require this and the provider you plan to use for us to approve.